ISSN 2319 – 5746 EISSN 2319 – 5754

Species

In silico Comparative Genomics of Treponema

Anjaneyulu k[1], Ashok .P. Patil[2], Desai PV[3*]

1. Ph.d Scholar, Dept. of Bio Sciences., South Gujarat University , Post Box No 49, Surat - 395007, Gujarat, INDIA
2. Ph.d Scholar, Dept. of Bio Sciences., South Gujarat University , Post Box No 49, Surat - 395007, Gujarat, INDIA
3. Professor , Dept. of Bio Sciences., South Gujarat University , Post Box No 49, Surat - 395007, Gujarat, INDIA

***Corresponding author:** Ph.d Scholar, Dept. of Bio Sciences, South Gujarat University, Post Box No 49, Surat - 395007, Gujarat, INDIA, E-Mail: anjaneyu.k@gmail.com

Received 04 August; accepted 18 September; published online 01 October; printed 16 October 2012

ABSTRACT

T. pallidum is one of the few human bacterial pathogens that have not been cultivated in vitro. This pathogens still remains the enigmatic pathogen, since few of its virulence factors have been identified and the pathogenesis of the disease is poorly understood. Several experimental approaches such as evolutionary or mutation analysis and complementation to definitively identify virulence determinants are in infancy state. Whole genome sequencing of the available *Treponema* subspecies and the resulting comparative analysis of genome sequences approaches seems to be one a promising approach in the whole context. Divergence within the species is mainly caused by variation in gene and protein sequences but also by differences in the set of genes that is present in a particular species. Proteins that are specific for a particular species may be responsible for its adapted phenotype, e.g. its ability to act as a pathogen or its resistance to a certain drug. Identifying species-specific proteins is thus a relevant aim, and here we make a small contribution towards its achievement. In the present work, we have compared the genomes, genes and proteins of five different Treponema sub species by extracting numerous protein sequence properties using state-of-the-art Support Vector Machines. The genome of the *Treponema pallidum* sub-species were sequenced to study the gene properties (Comparative Genome Sequencing, CGS) about 5016 protein coding genes. The sequences were compared with multiplewise alignment tool. The result obtained was filtered on the basis of sequences with 100 percent and 99 percent similarity. Functional sites of these sequences were predicted with the help of prosite scan. When compared to the heterogeneity in the *T. pallidum* chromosome. To our surprise, we find that proteins of different species are signicantly correlated and can be distinguished based on sequence properties and functional sites encoded in within their genomes. This discrimination does not rely on any homology criteria but is based only on the biophysical characteristics en-coded in the sequence. We have also constructed a phylogenetic tree based on the results of the comparisons, and compared it to the well-documented. The observed gene and protein comparison is the first assessment of the degree of variation between the five *T.pallidium* sub-species and hence it paves the way for phylogenetic studies of these enigmatic organisms. Moreover the divergence in genome, genes and proteins more often belonged to the group of genes with predicted virulence and unknown functions suggesting their involvement in infection differences such as yaws or syphilis.

Keywords: T.pallidium sub-species; Comparative Genomeics, Gene Wiz Browser, Prosite, Cladogram.

Abbreviations: GW – GeneWiz, Bp –Base Pairs, AA –Amino Acids, CGS - Comparative Genome Sequencing.

Homology:
The relationship among sequences due to descent from a common ancestral sequence. An important organizing principle for genomic studies because structural and functional similarities tend to change together along the structure of homology relationships.

1. INTRODUCTION

Both comparative gene analysis as well as comparative proteins encoded in complete genomes of an organism revels novel and unique species specific information inspite of the techniques are still being in their infancy, and have yet to reach their full potential. A comparative genomics is a powerful tool which enables to understand the underlying mechanism of evolution, pathogenesis and adaptive strategies in emerging non cultural pathogens such as Spirochetes. *T. pallidumi spp* is one of the few unusual human pathogens that have not been cultured continuously in vitro. A Gram-negative spirochaete bacterium with subspecies cause treponemal diseases such as syphilis, bejel, pinta and yaws. The treponemes have a cytoplasmic and outer membrane. Five subspecies of *Treponema pallidum* namely *Treponema pallidum subsps. pallidum DAL1, Treponema pallidum subsps. pallidum SS14, Treponema pallidum subsps. pallidum str., Treponema pallidum subsps. pallidum str. Nichols* and *Treponema pallidum subsps. pallidum str. CDC2* were compared on the basis of some genomic properties and various types of functional proteins. Closely related species reveal species-specific differences and evolutionary selection pressures on genes (Lukashin and Borodovsky, 1998). At the same time, a comparative sequence analysis provides the means for a better annotation. In addition to its spirochetal morphology and absence of lipopolysaccharide in its outer membrane contains relatively few intra membranous proteins; put forward several limitations on its research. With an objective to gain more insight in the functional elements of the genomes of the species of *Treponema* ,present study reveals the comparative relationship within the predicted ORFs , protein sequences with their functional sites , encoded in the complete genomes of *Treponema pallidum spp.* Conclusions from such Comparative analysis augment the understanding of the host-parasite interactions that enable pathogens to carve out unique ecological niches in nature. In the present study, we summarize the findings of each genome along with a computational comparative analysis of the five different subspecies of *T. pallidum* that can provide further insight into species and strain uniqueness and importantly can stimulate new studies leading to new approaches into disease prevention and treatment (Lowe and Eddy, 1997).

2. Statement of the Problem

The present comparative analysis among the *Treponema* subspecies which are pathogenic to human, aims to allocate the similarity and differences among the genes encoded genomes. The comparative study also aims to

SCIENTOMETRIC

Interpreting the functional content of a given genomic sequence is one of the central challenges of biology today. Perhaps the most promising approach to this problem is based on the comparative method of classic biology in the modern guise of sequence comparison. For instance, protein-coding regions tend to be conserved between species. Hence, a simple method for distinguishing a functional exon from the chance absence of stop codons is to investigate its homologue from closely related species.

Citation analysis:
It is the examination of the frequency, patterns, and graphs of citations in articles a nd books. It uses citations in scholarly works to establish links to other works or other researchers. Citation analysis is one of the most widely used methods of bibliometrics.

Forward genetics:
It involves studying genes one at a time. Only a small minority of genes are uniquely associated with an easily definable phenotype - a characteristic that is critical for determining gene function by forward genetics.

reveal the genomic properties of the species, identify the identical sequences of proteins and analyze their functional sites.

2.1 Scope of the Study

The aim of the study is to allocate the common and differentiating functional components encoded within the genomes of the five species of *Treponema*. Research review signifies the similarities and differences at morphological level as, *Treponema pallidum* subspecies are morphologically and serologically indistinguishable. The mode of transmission is not unique in nature. The course of each disease is significantly variable. The outer membrane of *T. pallidum* has too few surface proteins for an antibody to be effective. Thus due to poor antigencity, it's diagnosis and treatment through antibodies (vaccines) is difficult. The molecular analysis at sequence point such as on their mode of pathogencity and study the clinical significance of these species. Comparative genome, gene and protein analysis of the five subspecies of their findings the similarity and differences may be useful for future research

2.2 Limitations of the Study

- Study undertaken is limited to three years
- The genomes of the species Treponema was available during the timeline of this research study
- We did the citation analysis based on the Secondary information available in the databases
- In this study we did not include citiation analysis based on the invitro findings

3. MATERIALS AND METHODS

3 .1. Materials

3.1.1. Genome Sequences

Genomes of the *Treponema pallidum subsps* namely *Treponema pallidum subsps. pallidum DAL1* (species A), *Treponema pallidum subsps. pallidum SS14* (species B), *Treponema pallidum subsps. pallidum str. Chicago* (species C), *Treponema pallidum subsps. pallidum str. Nichols* (species D) and *Treponema pallidum subsps. pallidum str. CDC2* (species E) were selected for analysis and abbrivated as above. Genome sequences of all the *Treponema* subspecies and genome statistics were collected from the Genome sequence database maintained at the National center for Biotechnology Information (National Institutes of Health, Bethesda, Md.). This resourse organizes information on genomes including sequences, maps, chromosomes, assemblies and annotations (http://www.ncbi.nlm.nih. gov/sites/entrez?Db=genome).

3.2. Research Methodology

3.2.1. Sequence analysis: detection and interpretation of varying levels of genome sequence similarity

3.2.1.1. Clustal W Multiple Wise Alignments Program

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins. It attempts to calculate the best match for the selected sequences and lines them up so that the identities, similarities and differences can be seen (http://www.ebi.ac.uk/Tools/msa/clustalw2/#).

3.2.1.2. Tree View Software

Phylogenetic trees were constructed using the CLUSTALW programs (Sigrist et al., 2005) with the neighbor-joining and least squares (Fitch-Margoliash) methods, accompanied by bootstrap analysis (De Castro et al., 2005). Tree View is a program for displaying and printing phylogenies. The program reads most NEXUS tree files (such as those produced by PAUP and COMPONENT) and PHYLIP style tree files (including those produced by fast DNAml and CLUSTALW).

3.2.1.3. GeneWiz browser 0.94 server

GeneWiz browser 0.94 server is an interactive web application for visualizing genomic data of prokaryotic chromosomes. The tool allows users to carry out various analyses such as mapping alignments of homologous genes to other genomes, mapping of short sequencing reads to a reference chromosome and calculating DNA properties such as curvature or stac k-ing energy along the chromosome (Tamura et al., 2007). The GeneWiz browser produces an interactive graphic that enables zooming from a global scale down to single nucleotides without changing the size of the plot. Its ability to disproportionally zoom provides optimal readability and increased functionality compared to other browsers. It allows the user to select the display of various genomic features such as color setting and data ranges. Custom numerical data can be added to the plot allowing, for example, visualization of gene expression and regulation data. Further, standard atlases are pre-generated for all prokaryotic genomes available in GenBank, providing a fast overview of all available genomes, including recently deposited genome sequences. The tool is available online from (http://www.cbs.dtu.dk/services/gwBrowser).

3.2.1.4. Microbial Genome Annotation Tools

GLIMMER is a system for finding genes in microbial DNA, especially the genomes of bacteria and archaea. GLIMMER (Gene Locator and Interpolated Markov ModelER) uses interpolated Markov models to identify coding regions (elcher et al., 1999), (http://www.ncbi.nlm.nih.gov/genomes /MICROBES/glimmer_3.cgi?).

3.2.2. Conservation and diversity of functional classes of proteins between the subspecies of *Treponema*

Recent advances in high-throughput structural determination techniques and structural genomics initiatives have produced an increase in volume of structural data for proteins prior to knowledge of their functions. With these advances several tools are developed rapidly to predict functions for proteins based on their sequence similarity.

3.2.2.1. Prosite

PROSITE is a database of protein, currently contains patterns and profiles specific for more than a thousand protein families or domains. It is based on the observation that large number of different proteins can be grouped on the basis of similarities in their sequences, into a limited number of families. Proteins or protein domains belonging to a particular family generally share functional attributes and are derived from a common ancestor. The ProRule section of PROSITE is constituted of manually created rules that can automatically generate annotation in the UniProtKB/Swiss-Prot format based on PROSITE motifs. These rules, most of the times rules are based on PROSITE profiles as they are more specific than patterns, but occasionally rules make use of patterns. In these cases, the rules will not work independently, but will be called by another rule, which will be triggered by a profile. In addition to these rules corresponding to a unique PROSITE motif, there are also rules triggered by a specific combination of PROSITE motifs called metamotifs. Metamotifs allow the definition of arrangements of domains separated by spacers of variable size, as well as the anchoring to the N- and/or C-termini and the exclusion of a PROSITE motif (Sigrist et al., 2010). ProRule is used to create UniProtKB/Swiss-Prot lines with basic and complex annotation derived from the

TABLE 1 GENOME PROPERTIES

Species	Size (Mb)	GC%	Gene	Protein
A	1.14	52.8	1,119	1,056
B	1.14	52.8	1,088	1,028
C	1.14	52.8	1,118	981
D	1.14	52.8	1,095	1,036
E	1.14	52.8	1,122	1,065

presence of the domain and of biologically critical amino acids: domain name and boundaries, EC number, function, keywords, associated PROSITE patterns, PTMs, active sites, disulfide bonds, etc.). ProRule contains notably the position of structurally and/or functionally critical amino acid(s), as well as the condition(s) they must fulfil to play their biological role(s). Part of these supplementary data are used by ScanProsite that not only provides the protein sequence matched by a profile, but also information about the relevance of biologically meaningful residues, like active sites, binding sites, post-translational modification sites or disulfide bonds, to help function determination.

4. RESULTS AND DISCUSSION

4.1. Comparative Genome Analysis

Completely automated computational analysis of genome sequences of five subspecies of *Treponema pallidum* was obtained from NCBI to compare the basic properties of genes of these species. The size of the genome was found to be 1.4Mb for all species under analysis. Table 1 clearly indicated the result of the comparative analysis of the subspecies the content of GC% is almost same about 52.8%. Number of genes in species A, C and E were in the range of 1118 to 1122 whereas species B and D had comparatively differed to some extent. Number of proteins was almost same in number in species A and E respectively; and B and D. Species C showed less number of proteins counting only till 981 proteins.

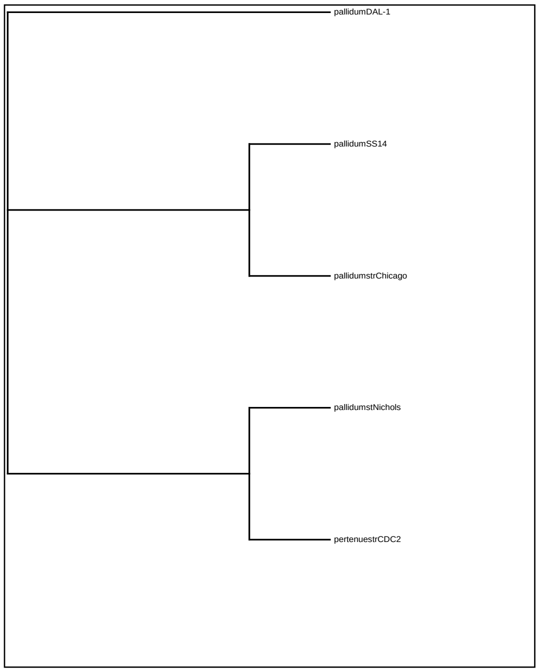


Figure 1

Phylogenetic Tree Depicting the Relationships between *T. Pallidum* subspecies

4.2. Tree View Software Analysis

In addition to the species discrimination, it was interesting to explore whether using sequence features to discriminate between bacterial subspecies by machine learning will provide an accurate phylogenetic relationship between the subspecies as documented in Fig 1.

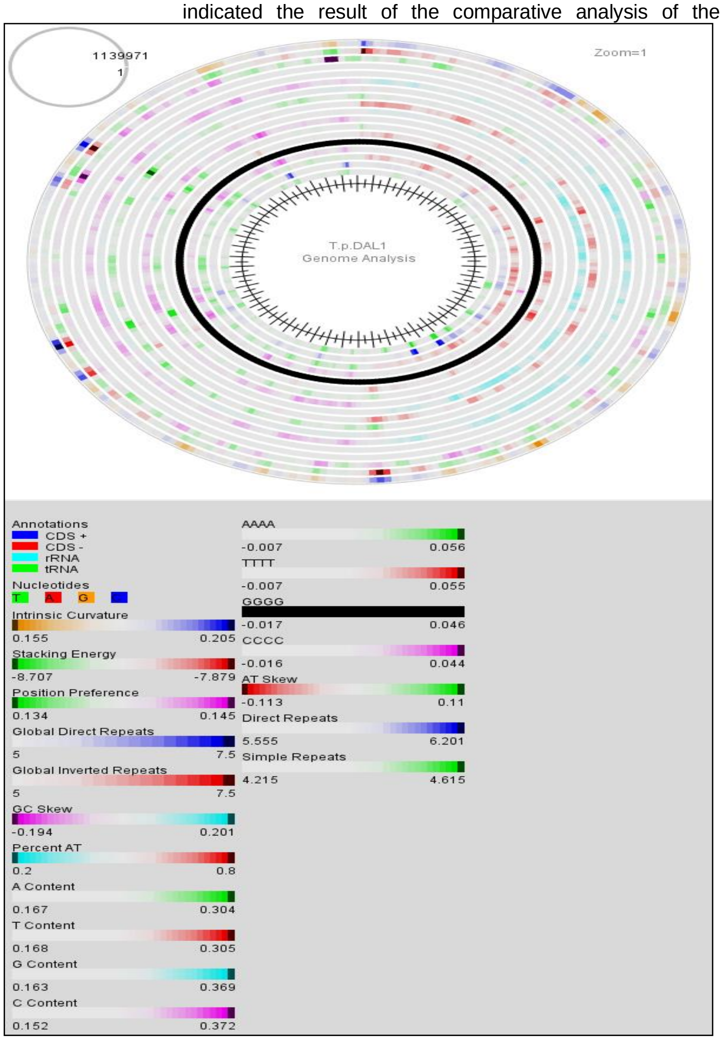


Figure 2

Genome map of Treponema palladium DAL1

4.3. Genewiz Browser Results

4.3.1. *Treponema pallidum* subsp. *pallidum DAL-1*

The Lineage: Bacteria - Spirochaetes - Spirochaetales - Spirochaetaceae; Treponema - Treponema pallidum - Treponema pallidum subsp. pallidum - Treponema pallidum subsp. pallidum DAL1.

Treponema pallidum subsp. *pallidum* DAL1: This organism is the causative agent of endemic and venereal syphilis. This sexual transmitted disease was first discovered in Europe at the end of the fifteenth century, however, the causative agent was not identified until 1905. At one time syphilis was the third most commonly reported communicable disease in the USA. Syphilis is characterized by multiple clinical stages and long periods of latent, asymptomatic infection. Although effective therapies have been available since the introduction of penicillin, syphilis remains a global health problem. *Treponema pallidum*subsp. *pallidum* str. Dallas1. This strain will be used for comparative analysis, Fig.2 shows the Genome map of Treponema palladium DAL1.

- Lane 1 = feature lane (annotations),
- Lane 2 = nucleotides

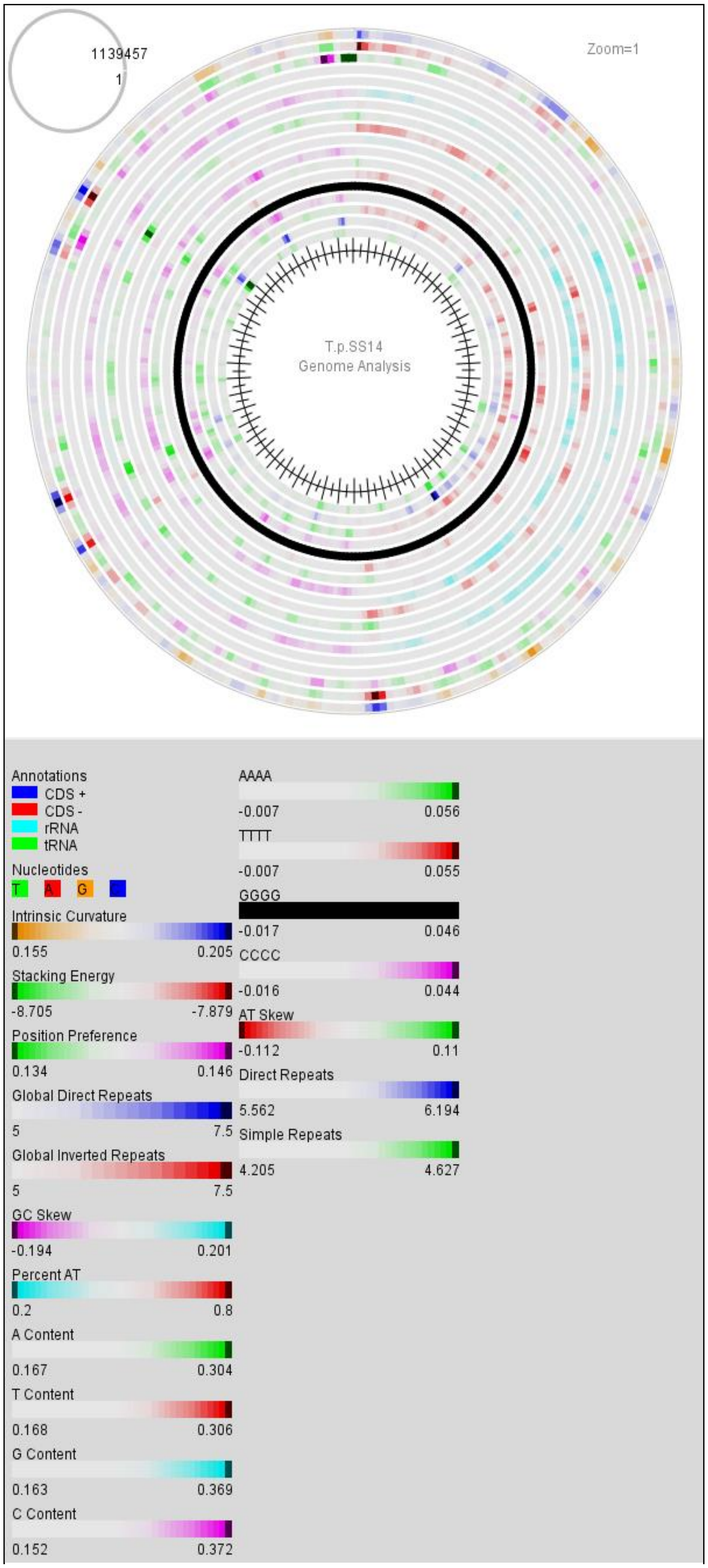


Figure 3

Genome map of *Treponema palladium SS14*

- Lane 3 = intrinsic curvature
- Lane 4 = stacking energy
- Lane 5 = positional preferences
- Lanes 6 and 7 = Global direct repeats and global inverted repeats
- Lane 8 = GC skew
- Lane 9 = percent AT
- Lanes 10, 11, 12 and 13 = A, T, G and C content respectively
- Lanes 14, 15, 16 and 17= AAAA, TTTT, GGGG and CCCC repeats respectively
- Lane 18 = AT skew
- Lanes 19 and 20 = direct repeats and simple repeats

Genes in lines are color-coded according to the following category:

- Wine Red = The genes involved in central metabolism and respiration without orthologues in H.pyloricyan, methyl-accepting chemotaxis proteins (MCPs)
- Dark Blue = Type IV secretion system
- Sky Blue = Genes involved in acid acclimation
- Green = Putative secreted virulence factors
- Pale Green = Glycosyltransferse gene cluster specific of H.bizzozeronii;

Pale Grey = All other CDSs. ACC, acetophenone carboxylase; comB, Type IV secretion system; NAP, periplasmic nitrate reductase; AHD, allophanate hydrolase; GT, glycosyltransferase; NRS, nitrite reductase system; SNO, S and N oxidases; FDH, formate reductase system; PL, polysaccharide lyase

4.3.2. Treponema pallidum subsp. pallidum SS14

The Lineage: Bacteria - Spirochaetes - Spirochaetales - Spirochaetaceae; Treponema - Treponema pallidum - Treponema pallidum subsp. pallidum - Treponema pallidum subsp. pallidum SS14.

Treponema pallidum subsp. *pallidum SS14*: This organism is the causative agent of endemic and venereal syphilis. This sexual transmitted disease was first discovered in Europe at the end of the fifteenth century; however, the causative agent was not identified until 1905. At one time syphilis was the third most commonly reported communicable disease in the USA. Syphilis is characterized by multiple clinical stages and long periods of latent, asymptomatic infection. Although effective therapies have been available since the introduction of penicillin, syphilis remains a global health problem. *Treponema pallidum* subsp. *pallidum* SS14. *Treponemapallidum* subsp. *pallidum* SS14 was isolated in 1977 from a patient with secondary syphilis. This strain is less susceptible than the Nichols strain for a number of antibiotics and will be used for comparative analysis. Fig.3 shows the Genome map of Treponema palladium SS14.

4.3.3. Treponema pallidum subsp. pallidum str. Chicago

The Lineage: Bacteria - Spirochaetes - Spirochaetales - Spirochaetaceae; Treponema - Treponema pallidum - Treponema pallidum subsp. pallidum - Treponema pallidum subsp. pallidum str. Chicago.

Treponema pallidum subsp. *pallidum* str. Chicago: The availability of more *Treponema pallidum* genomes will greatly help comparative studies among isolates; facilitate the improvement of typing methods and the identification of potential targets to be used as protective antigens. Fig. 4 shows the genome map of *Treponema palladium str.* Chicago.

4.3.4. Treponema pallidum subsp. pallidum str. Nichols

The Lineage: Bacteria - Spirochaetes - Spirochaetales - Spirochaetaceae; Treponema - Treponema pallidum - Treponema pallidum subsp. pallidum - Treponema pallidum subsp. pallidum str. Nichols.

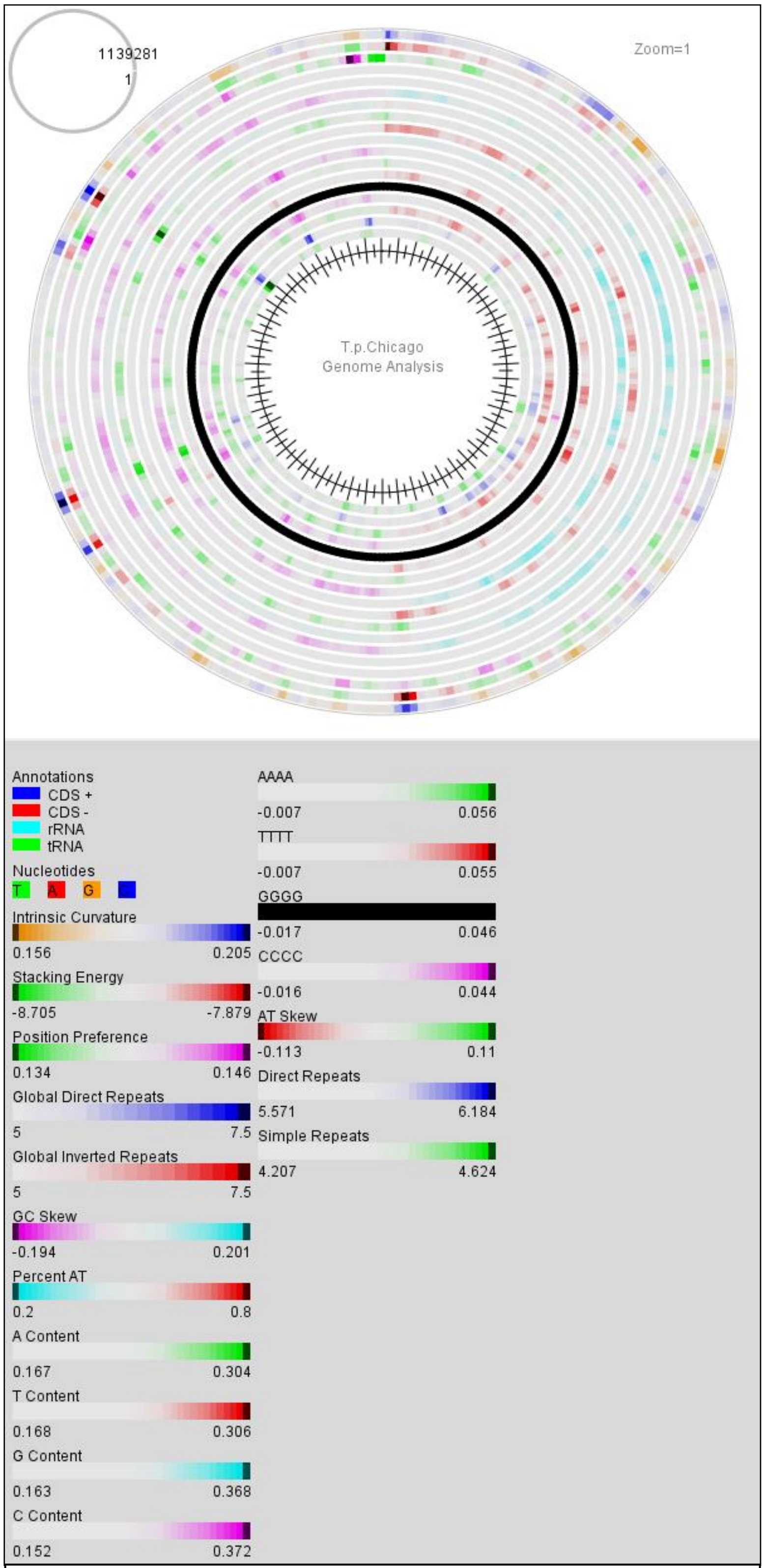


Figure 4

Genome map of *Treponema palladium str.* Chicago

Treponema pallidum subsp. *pallidum*: This organism is the causative agent of endemic and venereal syphilis. This sexual transmitted disease was first discovered in Europe at the end of the fifteenth century, however, the causative agent was not identified until 1905. At one time syphilis was the third most commonly reported communicable disease in the USA. Syphilis is characterized by multiple clinical stages and long periods of latent, asymptomatic infection. Although effective therapies have been available since the introduction of penicillin, syphilis remains a global health problem. *Treponema pallidum* subsp. *pallidum* strain Nichols, this strain was originally isolated in 1912 from a neurosyphilitic patient and is virulent. Fig.5 shows the genome map of *Treponema palladium str.* Nichols

4.3.5. Treponema pallidum subsp. pertenue str. CDC2

The Lineage: Bacteria - Spirochaetes - Spirochaetales - Spirochaetaceae; Treponema - Treponema pallidum - Treponema pallidum subsp. pallidum - Treponema pallidum subsp. pallidum str. CDC2.

Treponema pallidum subsp. *pertenue*: This species causes chronic and disfiguring illness called yaws. The disease starts as a skin infection causing persistent ulcers and progresses to form tumor-like masses. This disease tends to infect children and is common in rural areas in Africa, Southeast Asia and equatorial South America. *Treponema pallidum* subsp. *pertenue* str. CDC2, this strain was isolated in Akorabo, Ghana in 1980 and will be used for comparative analysis. Fig. 6 shows the genome map of *Treponema palladium str. CDC2.*

The collective analysis of the each of the genome characterization attained from the Genewiz Browser summarized in the Table 2. This table illustrates the comparative results obtained from Genewiz browser to study the DNA characteristics of genomes. All the DNA properties of the five subspecies were found to be identical expect the direct repeats and inverted repeats which distinguished them from each other. Direct repeats were similar in species B and E whereas other three species had difference in number. Inverted repeats were identical in species A and E while B and C showed a minute difference in number.

4.4. Comparative analysis of the Proteins present in Treponema Species

Table 2 shows the Comparative genome Analysis and Properties

4.4.1. Protein Sequence Alignment analysis

4.4.1.1. Protein sequence with 100% Similarity

The protein sequences were obtained from NCBI genome browser for each *Treponema pallidum* subspecies. About 5061 sequences were compared to each other. Multiple sequence alignment was executed by using ClustalW software. The table 3 denotes total number of sequences of five species having 100% similarity based on related type of proteins. The analysis performed, resulted into 92 sequences of these five species which showed 100% similarity when matched with each other. It was observed that species D has the highest number of 43 sequences matched with other four species.Species C and D have the maximum 100% score alignment of 13 sequences while species D and E and have 10 aligned similar sequences.Pairing between species A and E; and B and D were found to be having 10 sequences with complete similar protein based sequences. Species B and E showed the least number of 6 sequences aligned score of 100. Table 3 shows the Total number of protein Sequences with aligned score of 100% of five *Treponema pallidum subspecies.*

4.4.1.2. Protein sequence with 99% Similarity

The protein sequences were obtained from NCBI genome browser for each *Treponema pallidum* subspecies. About 5061 sequences were compared to each other. Multiple sequence alignment was executed by using ClustalW

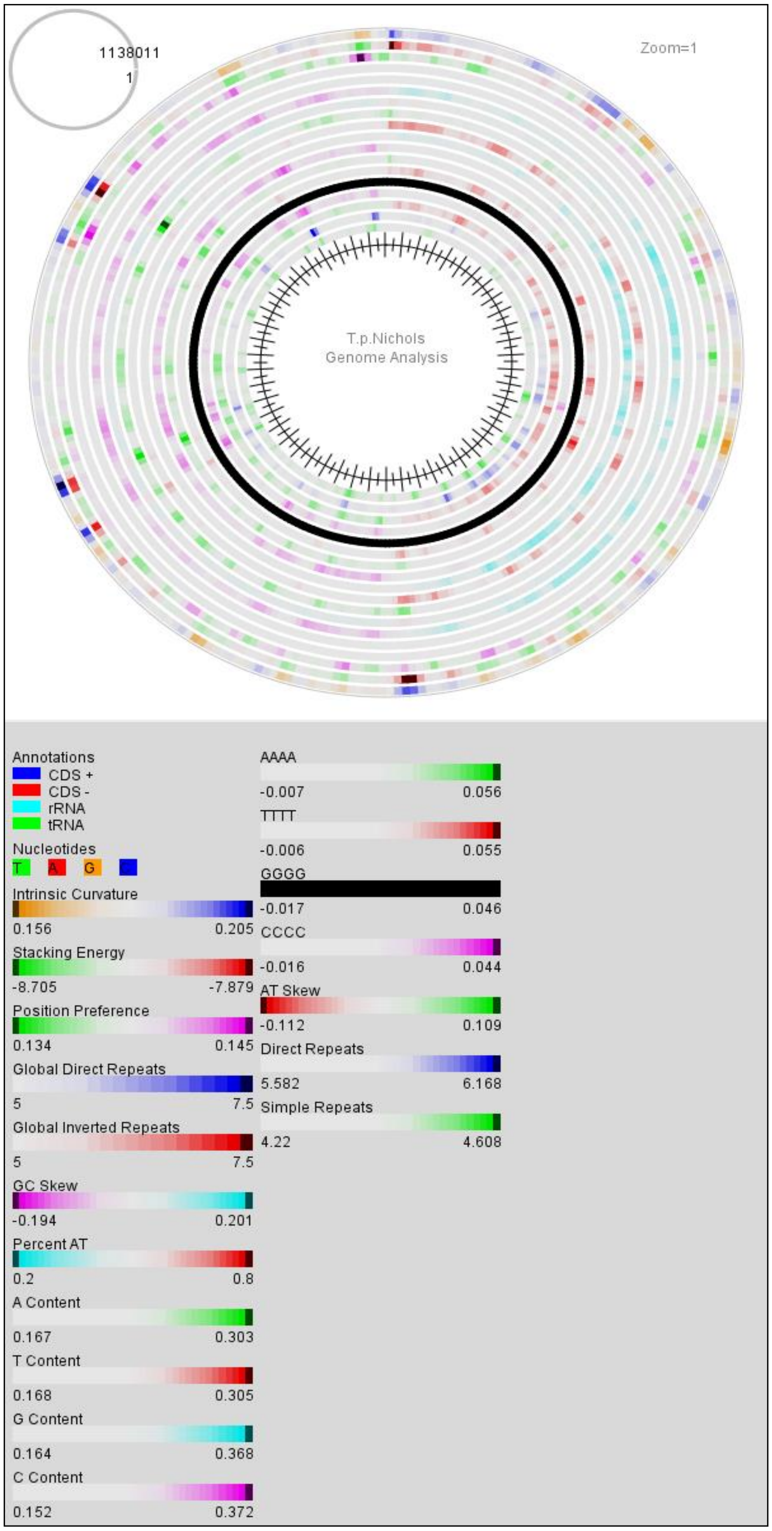


Figure 5

Genome map of *Treponema palladium str*. Nichols

software. The above table denotes total number of sequences of five species having 99% similarity based on related type of proteins. The analysis resulted into 13 sequences of these five species which showed 99% similarity when matched with each other. It was observed that Species D and E have the most 99% similar sequences about 5 sequences. Species C and E have 4 sequences with 99% similar sequences. Table 4 shows the total number of protein Sequences with aligned score of 99% of five Treponema pallidum subspecies

4.4.2. Comparative analysis of Protein Based on functional categories

4.4.2.1. Distribution of Proteins (100% Similarity) based on Functional categories

Using ClustalW software, 92 protein sequences were filtered based on sequences having 100% similarity. Out of 92 sequences, 26 types of different proteins were categorized. The above table details about the presence of a specific type of protein in an single subspecies among 92 sequences having 100% similarity.Analysis reveals that among all similar proteins, ribosomal proteins L15 and L30 and Replication initiator factor proteins were most common to all the five subspecies of *Treponema pallidum*. Aspartyl glutamyl / tRNA amidotransferase subunit C and hypothetical proteins were the other two types of proteins commonly found in all five subspecies of Treponema. Special types of putative proteins were found in all five species with different functional proteins. Aspartyl glutamyl /tRNA amidotransferase subunit A proteins and lipoproteins were observed in all four species except species E. Protein like methionine aminopeptidase was found in species B, C and D but not in A and E. Phosphoenol pyruvate carboxykinase wa found in species C, D and E except Species A and B respectively Table 5 shows the Functional categories within the 100% similar protein sequences in *Treponema pallidum* subspecies ('#' indicates the presence of hypothetical proteins with other type of protein according to the databases).

4.4.2.2. Distribution of proteins (99% similar) based on Functional categories

Using ClustalW software, 13 protein sequences were filtered based on sequences having 99% similarity. Out of 13 sequences, 10 types of different proteins were categorized. The above table details about the presence of a specific type of protein in a single subspecies among 13 sequences having 100% similarity. Table 6 shows the Comparative analysis of based on proteins present in *Treponema pallidum* subspecies with 99% similarity sequences. Species C and D had shown maximum similarity in Apolipoprotein N-acyltransferase protein and Alginate O-acetylation protein (algl). Species A and B have Spermidine/putrescine ABC superfamily ATP binding cassette transporter, ABC protein and Species C and E have 30S ribosomal protein S9. Table 6 shows the comparative analysis of based on proteins present in *Treponema pallidum* subspecies with 99% similarity sequences. (#' indicates the presence of hypothetical proteins with other type of protein according to the databases).

4.5. Protein Functional Site Analysis

It is apparent, when studying protein sequence families, that some regions have been better conserved than others during evolution. These conserved regions are generally important for the three dimensional structure and function of a protein. By analyzing the constant and variable properties of such groups of similar sequences, it is possible to derive a signature for a protein family or domain, which distinguishes its members from all other unrelated proteins. A significant analogy is to use the fingerprints for identification. A fingerprint, a protein signature can be used to assign a new protein to a specific family of proteins and thus to formulate hypotheses about its function.

4.5.1. Comparative Functional Sites in Proteins (100% Similar)

92 proteins scanned with the prosite for predictind the functional sites and the locations. Out of these we have got 28 functional hits. Majorly found are NHL repeat proteins, recombinase A protein and Spermidine/putrescine ABC superfamily ATP binding cassette transporter, ABC protein

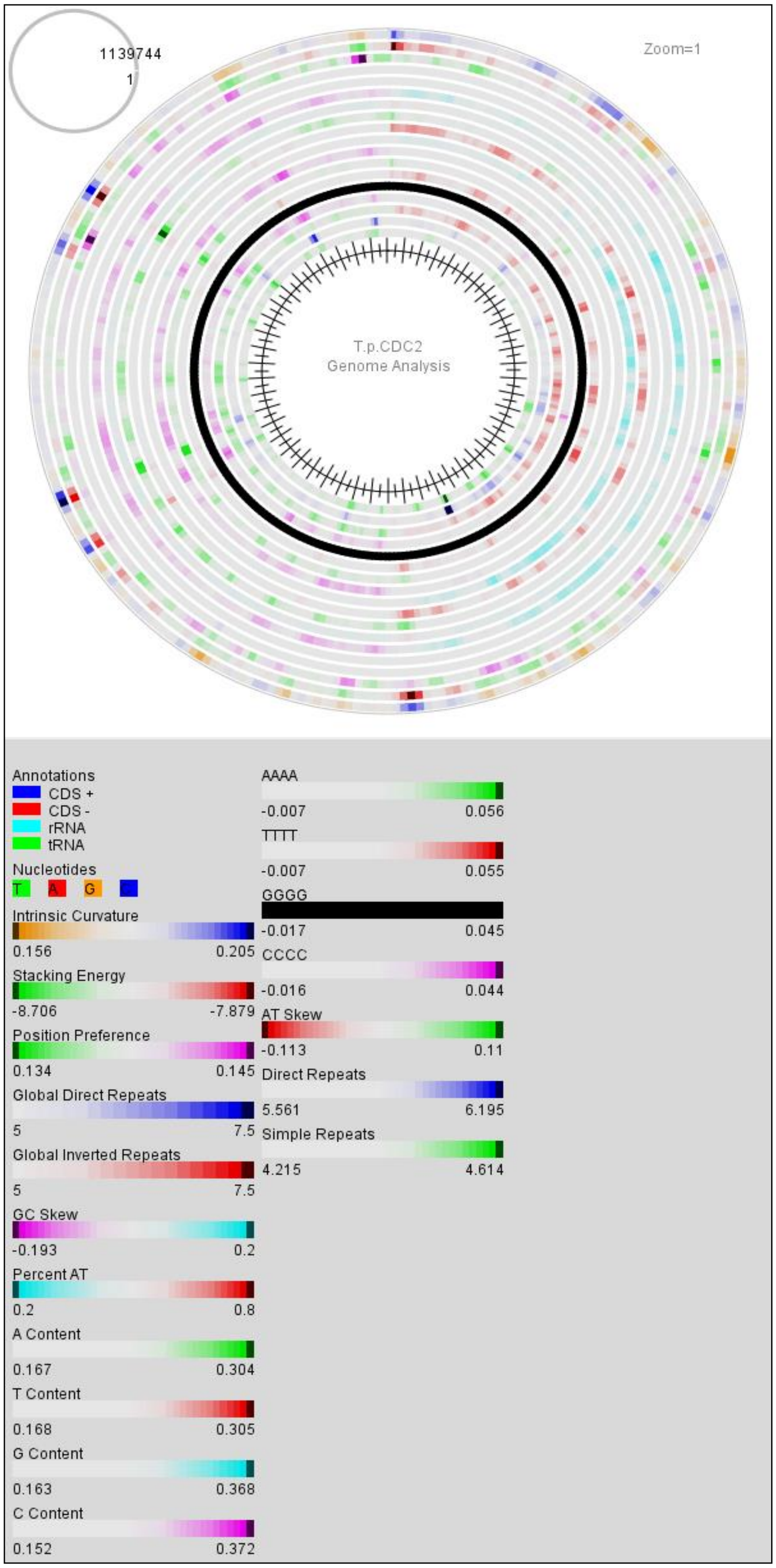
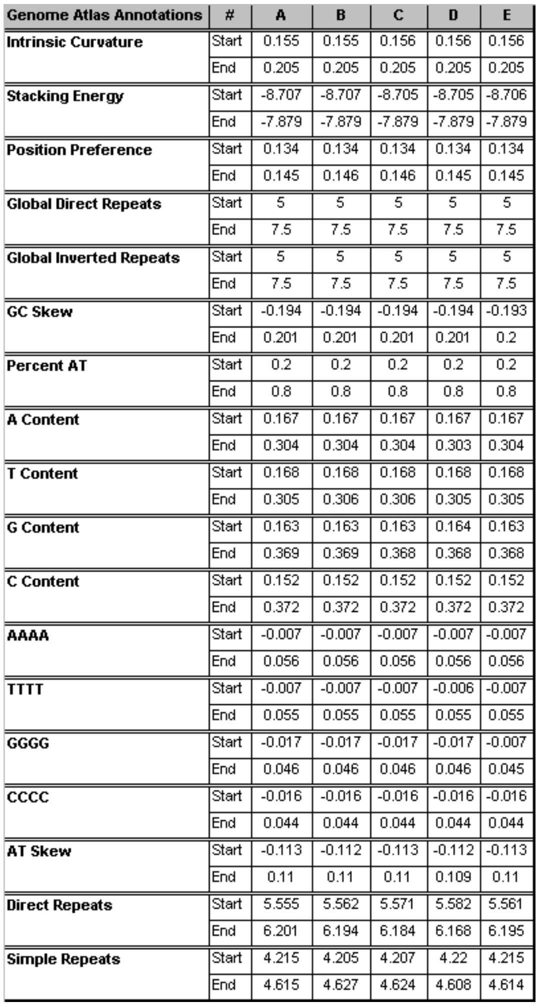


Figure 6

Genome map of *Treponema palladium str.* CDC2

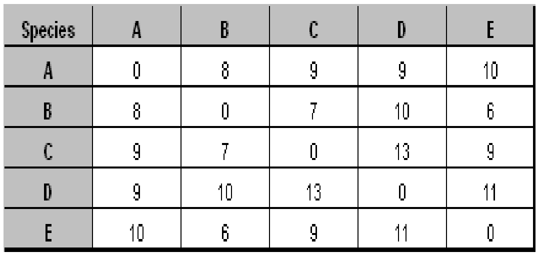
respectively. Table 7 shows the Predicted Functional site in total of 92 proteins

Table 2 Comparative Genome Analysis



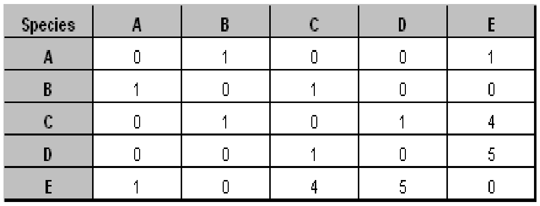
Genome Atlas Annotations	#	A	B	C	D	E
Intrinsic Curvature	Start	0.155	0.155	0.156	0.156	0.156
	End	0.205	0.205	0.205	0.205	0.205
Stacking Energy	Start	-8.707	-8.707	-8.705	-8.705	-8.706
	End	-7.879	-7.879	-7.879	-7.879	-7.879
Position Preference	Start	0.134	0.134	0.134	0.134	0.134
	End	0.145	0.146	0.146	0.145	0.145
Global Direct Repeats	Start	5	5	5	5	5
	End	7.5	7.5	7.5	7.5	7.5
Global Inverted Repeats	Start	5	5	5	5	5
	End	7.5	7.5	7.5	7.5	7.5
GC Skew	Start	-0.194	-0.194	-0.194	-0.194	-0.193
	End	0.201	0.201	0.201	0.201	0.2
Percent AT	Start	0.2	0.2	0.2	0.2	0.2
	End	0.8	0.8	0.8	0.8	0.8
A Content	Start	0.167	0.167	0.167	0.167	0.167
	End	0.304	0.304	0.304	0.303	0.304
T Content	Start	0.168	0.168	0.168	0.168	0.168
	End	0.305	0.306	0.306	0.305	0.305
G Content	Start	0.163	0.163	0.163	0.164	0.163
	End	0.369	0.369	0.368	0.368	0.368
C Content	Start	0.152	0.152	0.152	0.152	0.152
	End	0.372	0.372	0.372	0.372	0.372
AAAA	Start	-0.007	-0.007	-0.007	-0.007	-0.007
	End	0.056	0.056	0.056	0.056	0.056
TTTT	Start	-0.007	-0.007	-0.007	-0.006	-0.007
	End	0.055	0.055	0.055	0.055	0.055
GGGG	Start	-0.017	-0.017	-0.017	-0.017	-0.007
	End	0.046	0.046	0.046	0.046	0.045
CCCC	Start	-0.016	-0.016	-0.016	-0.016	-0.016
	End	0.044	0.044	0.044	0.044	0.044
AT Skew	Start	-0.113	-0.112	-0.113	-0.112	-0.113
	End	0.11	0.11	0.11	0.109	0.11
Direct Repeats	Start	5.555	5.562	5.571	5.582	5.561
	End	6.201	6.194	6.184	6.168	6.195
Simple Repeats	Start	4.215	4.205	4.207	4.22	4.215
	End	4.615	4.627	4.624	4.608	4.614

Table 3 Total number of protein Sequences with aligned score of 100% of five*Treponema pallidum*



Species	A	B	C	D	E
A	0	8	9	9	10
B	8	0	7	10	6
C	9	7	0	13	9
D	9	10	13	0	11
E	10	6	9	11	0

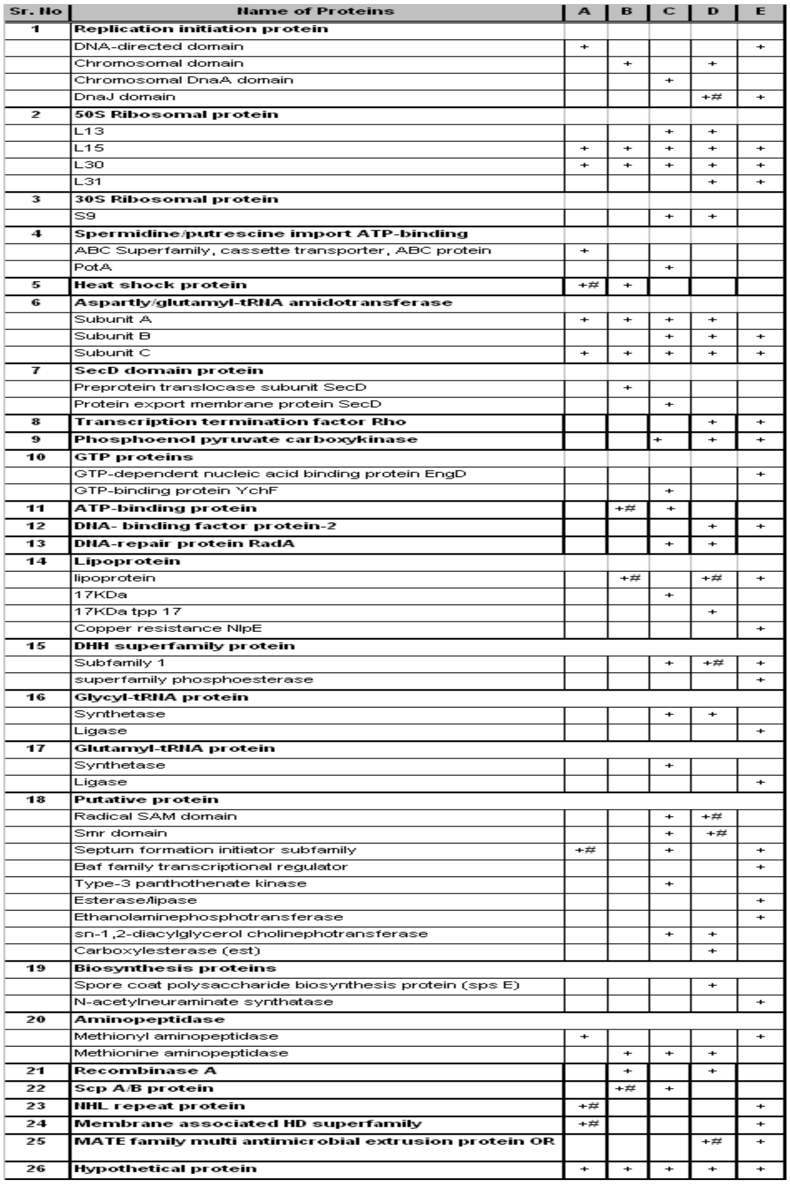
Table 4 Total number of protein Sequences with aligned score of 99% of five *Treponema pallidum* subspecies



Species	A	B	C	D	E
A	0	1	0	0	1
B	1	0	1	0	0
C	0	1	0	1	4
D	0	0	1	0	5
E	1	0	4	5	0

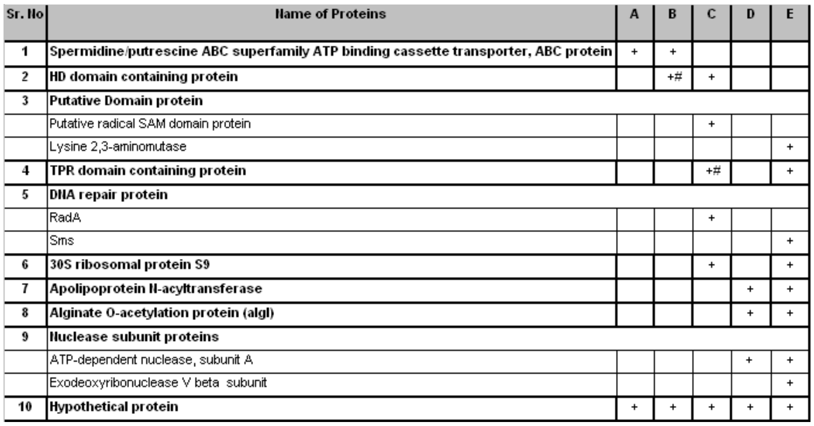
4.5.2. Comparative Functional Sites in Proteins (99% Similar)

Table 5 Functional categories within the 100% similar protein sequences in Treponema pallidum subspecies



Sr. No	Name of Proteins	A	B	C	D	E
1	**Replication initiation protein**					
	DNA-directed domain	+				+
	Chromosomal domain		+		+	
	Chromosomal DnaA domain			+		
	DnaJ domain				+#	+
2	**50S Ribosomal protein**					
	L13			+	+	
	L15	+	+	+	+	+
	L30	+	+	+	+	+
	L31				+	+
3	**30S Ribosomal protein**					
	S9			+	+	
4	**Spermidine/putrescine import ATP-binding**					
	ABC Superfamily, cassette transporter, ABC protein	+				
	PotA			+		
5	**Heat shock protein**	+#	+			
6	**Aspartly/glutamyl-tRNA amidotransferase**					
	Subunit A	+	+	+	+	
	Subunit B			+	+	+
	Subunit C	+	+	+	+	+
7	**SecD domain protein**					
	Preprotein translocase subunit SecD		+			
	Protein export membrane protein SecD			+		
8	**Transcription termination factor Rho**				+	+
9	**Phosphoenol pyruvate carboxykinase**			+	+	+
10	**GTP proteins**					
	GTP-dependent nucleic acid binding protein EngD					+
	GTP-binding protein YchF			+		
11	**ATP-binding protein**		+#	+		
12	**DNA- binding factor protein-2**				+	+
13	**DNA-repair protein RadA**			+	+	
14	**Lipoprotein**					
	lipoprotein		+#		+#	+
	17KDa			+		
	17KDa tpp 17				+	
	Copper resistance NlpE					+
15	**DHH superfamily protein**					
	Subfamily 1			+	+#	+
	superfamily phosphoesterase					+
16	**Glycyl-tRNA protein**					
	Synthetase			+	+	
	Ligase					+
17	**Glutamyl-tRNA protein**					
	Synthetase			+		
	Ligase					+
18	**Putative protein**					
	Radical SAM domain			+	+#	
	Smr domain			+	+#	
	Septum formation initiator subfamily	+#		+		+
	Baf family transcriptional regulator					+
	Type-3 panthothenate kinase			+		
	Esterase/lipase					+
	Ethanolaminephosphotransferase					+
	sn-1,2-diacylglycerol cholinephotransferase			+	+	
	Carboxylesterase (est)				+	
19	**Biosynthesis proteins**					
	Spore coat polysaccharide biosynthesis protein (sps E)				+	
	N-acetylneuraminate synthatase					+
20	**Aminopeptidase**					
	Methionyl aminopeptidase	+				+
	Methionine aminopeptidase		+	+	+	
21	**Recombinase A**		+		+	
22	**Scp A/B protein**		+#	+		
23	**NHL repeat protein**	+#				+
24	**Membrane associated HD superfamily**	+#				+
25	**MATE family multi antimicrobial extrusion protein OR**				+#	+
26	**Hypothetical protein**	+	+	+	+	+

Table 6 Comparative analysis of based on proteins present in Treponema pallidum subspecies with 99% similarity sequences.



Sr. No	Name of Proteins	A	B	C	D	E
1	**Spermidine/putrescine ABC superfamily ATP binding cassette transporter, ABC protein**	+	+			
2	**HD domain containing protein**		+#	+		
3	**Putative Domain protein**					
	Putative radical SAM domain protein			+		
	Lysine 2,3-aminomutase					+
4	**TPR domain containing protein**			+#		+
5	**DNA repair protein**					
	RadA			+		
	Sms					+
6	**30S ribosomal protein S9**			+		+
7	**Apolipoprotein N-acyltransferase**				+	+
8	**Alginate O-acetylation protein (algI)**				+	+
9	**Nuclease subunit proteins**					
	ATP-dependent nuclease, subunit A				+	+
	Exodeoxyribonuclease V beta subunit					+
10	**Hypothetical protein**	+	+	+	+	+

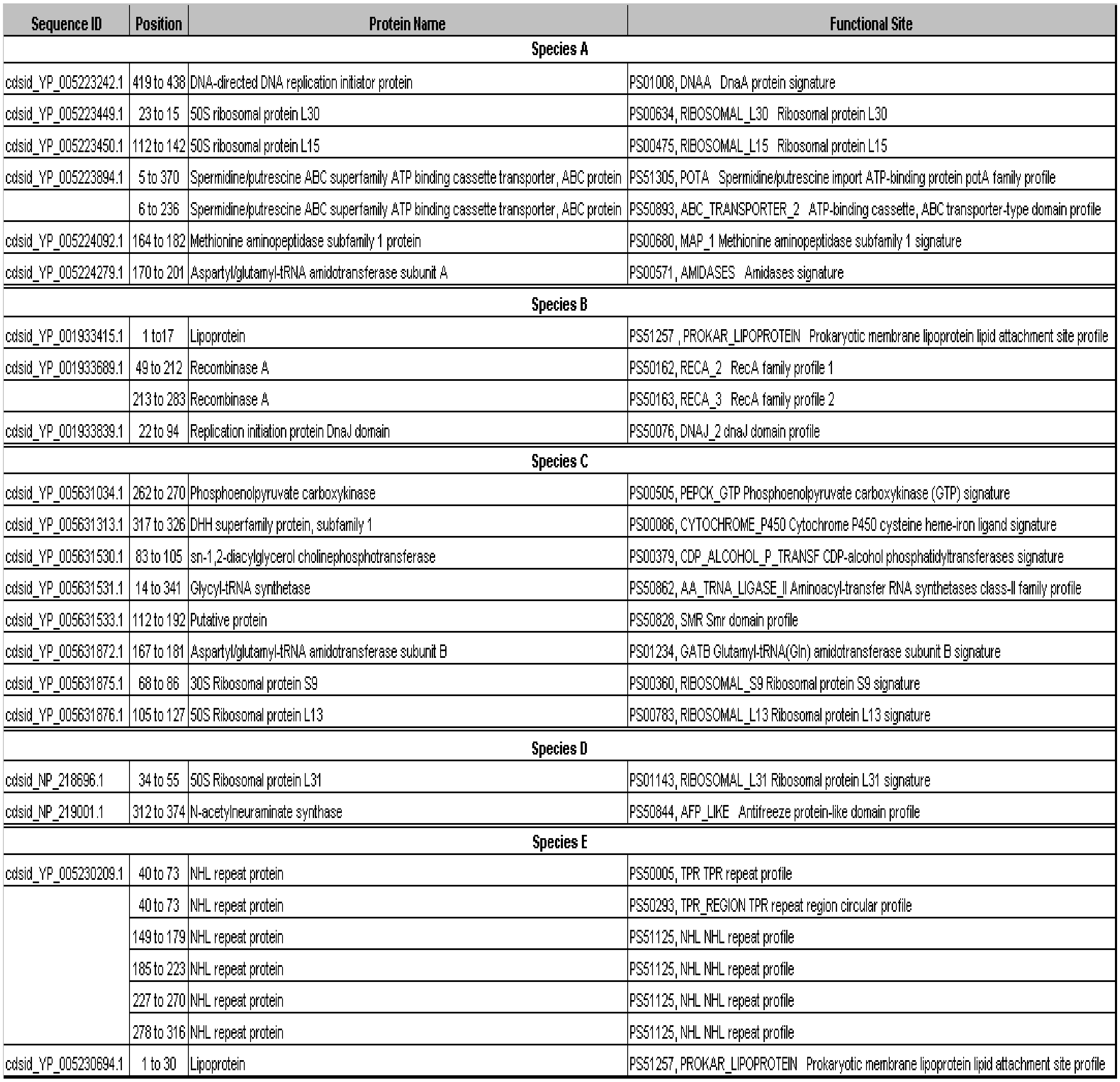
(#' indicates the presence of hypothetical proteins with other type of protein

13 proteins scanned with the prosite for predicting the functional sites and the locations. Out of these we have got 7 functional hits. Majorly found are ATP-binding cassette, ABC transporter-type domain and ATP- dependent nuclease, subunit A. Table 8: Predicted Functional site in total of 13 proteins.

5. CONCLUSION

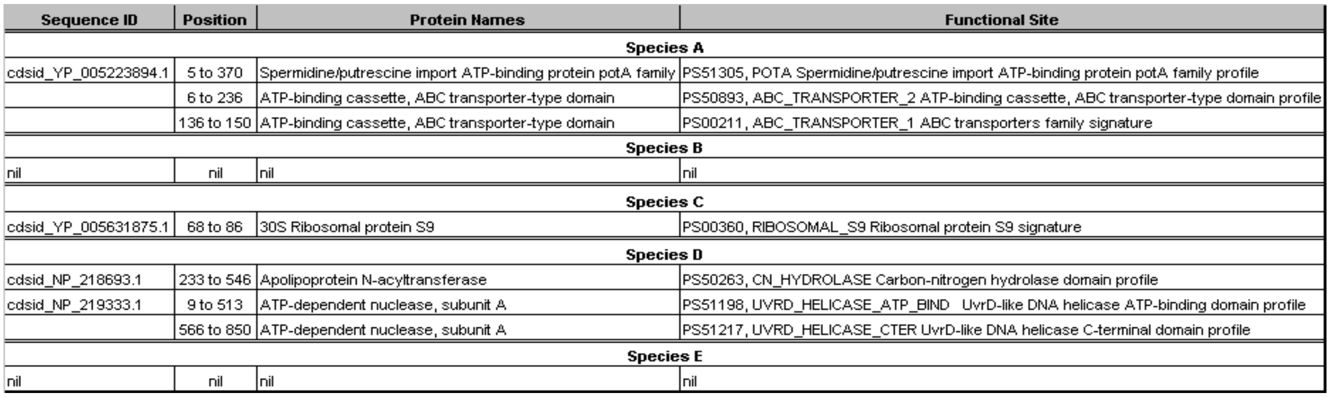
The comparative analysis of five subspecies of *Treponema pallidum* was performed to study their similarity and differences on the basis of comparison between their genomic properties and various types of proteins. With Bioinformatics tools and software used for analysis we are enable to conclude the differentiating characters among the subspecies of the *Treponema*. The genome sequences of the five subspecies of *T.pallidum* were extracted from NCBI namely Treponema pallidum DAL1 (species A), Treponema pallidum SS14 (species B), *Treponem pallidum str. Chicago* (species C), *Treponema pallidum str. Nichols* (species D) and *Treponema pallidum str. CDC2* (species E). Analysis of genomes from species A, B, C, D and E were performed on the basis of comparison with the genome sequences obtained from NCBI about 5166 sequences. The genomic properties and various types of proteins with their functional site were studied and compared among the five species to collect the information on their similarity and differences. Multiple sequence alignment was performed using Clustal W software of 5166 * 5166 sequences. The sequences aligned were filtered with the sequences having 100% and 99 % alignment scores based on the similar proteins present in the five subspecies. It resulted into 92 sequences with 100% identical protein sequences and 13 sequences with 99% identical proteins respectively. The functional site were obtained using Prositescan tool. 28 functional hits of 100% and 7 of 99% identical protein sequences were found to have similar functions. The detailed study and research are in mentioned tables for better interpretations of results and discussion. According to Table 1, we can infer that species B and D have similar number of genes and proteins whereas species A, C and E show similarity enomic properties. From Table 2; it can be observed that all DNA properties are similar in all five species. Species B and E show similar gradation in Direct repeats wherein species A and E and species B and C have identical Inverted repeats. Multiple sequence alignment using ClustalW software, screened all the five subspecies sequences and hence resulted that species C and D have 13 identical sequences of functional proteins with 100% alignment score and species D and E have 5 identical sequences of 99% similarity. The details are mentioned in Table 3 to 6 respectively. Table 7 and 8 gives information about the functional site of 28 proteins (100% identical) and 7 proteins (99% identical). The most analogous proteins are NHL repeat proteins, recombinase A protein and Spermidine/putrescine ABC superfamily ATP binding cassette transporter, ABC protein with 100% identical sequences whereas ATP-binding cassette, ABC transporter-type domain and ATP- dependent nuclease, subunit A with 99% identical sequences. Comparative genomics analysis between the species revealed that species B and D and species A and E are closely related to each other in their genomic composition while species C, D and E are similar in functional protein content. By means of local sequence similarity searches, Protein profile searches, and analysis of 100% and 99% similar protein funcational categories we have conducted a detailed comparative anal-ysis of the genomes of the *T.pallidum*. The level of conservation between functional classes and evolutionary measure, it was possible to characterize, in functional terms, the nature of the divergence between the five spirochetes and the common and distinct aspects of their physiological strategies.

Table 7 Predicted Functional site in total of 92 proteins



Sequence ID	Position	Protein Name	Functional Site
		Species A	
cdsid_YP_005223242.1	419 to 438	DNA-directed DNA replication initiator protein	PS01008, DNAA DnaA protein signature
cdsid_YP_005223449.1	23 to 15	50S ribosomal protein L30	PS00634, RIBOSOMAL_L30 Ribosomal protein L30
cdsid_YP_005223450.1	112 to 142	50S ribosomal protein L15	PS00475, RIBOSOMAL_L15 Ribosomal protein L15
cdsid_YP_005223894.1	5 to 370	Spermidine/putrescine ABC superfamily ATP binding cassette transporter, ABC protein	PS51305, POTA Spermidine/putrescine import ATP-binding protein potA family profile
	6 to 236	Spermidine/putrescine ABC superfamily ATP binding cassette transporter, ABC protein	PS50893, ABC_TRANSPORTER_2 ATP-binding cassette, ABC transporter-type domain profile
cdsid_YP_005224092.1	164 to 182	Methionine aminopeptidase subfamily 1 protein	PS00680, MAP_1 Methionine aminopeptidase subfamily 1 signature
cdsid_YP_005224279.1	170 to 201	Aspartyl/glutamyl-tRNA amidotransferase subunit A	PS00571, AMIDASES Amidases signature
		Species B	
cdsid_YP_001933415.1	1 to17	Lipoprotein	PS51257 , PROKAR_LIPOPROTEIN Prokaryotic membrane lipoprotein lipid attachment site profile
cdsid_YP_001933689.1	49 to 212	Recombinase A	PS50162, RECA_2 RecA family profile 1
	213 to 283	Recombinase A	PS50163, RECA_3 RecA family profile 2
cdsid_YP_001933839.1	22 to 94	Replication initiation protein DnaJ domain	PS50076, DNAJ_2 dnaJ domain profile
		Species C	
cdsid_YP_005631034.1	262 to 270	Phosphoenolpyruvate carboxykinase	PS00505, PEPCK_GTP Phosphoenolpyruvate carboxykinase (GTP) signature
cdsid_YP_005631313.1	317 to 326	DHH superfamily protein, subfamily 1	PS00086, CYTOCHROME_P450 Cytochrome P450 cysteine heme-iron ligand signature
cdsid_YP_005631530.1	83 to 105	sn-1,2-diacylglycerol cholinephosphotransferase	PS00379, CDP_ALCOHOL_P_TRANSF CDP-alcohol phosphatidyltransferases signature
cdsid_YP_005631531.1	14 to 341	Glycyl-tRNA synthetase	PS50862, AA_TRNA_LIGASE_II Aminoacyl-transfer RNA synthetases class-II family profile
cdsid_YP_005631533.1	112 to 192	Putative protein	PS50828, SMR Smr domain profile
cdsid_YP_005631872.1	167 to 181	Aspartyl/glutamyl-tRNA amidotransferase subunit B	PS01234, GATB Glutamyl-tRNA(Gln) amidotransferase subunit B signature
cdsid_YP_005631875.1	68 to 86	30S Ribosomal protein S9	PS00360, RIBOSOMAL_S9 Ribosomal protein S9 signature
cdsid_YP_005631876.1	105 to 127	50S Ribosomal protein L13	PS00783, RIBOSOMAL_L13 Ribosomal protein L13 signature
		Species D	
cdsid_NP_218696.1	34 to 55	50S Ribosomal protein L31	PS01143, RIBOSOMAL_L31 Ribosomal protein L31 signature
cdsid_NP_219001.1	312 to 374	N-acetylneuraminate synthase	PS50844, AFP_LIKE Antifreeze protein-like domain profile
		Species E	
cdsid_YP_005230209.1	40 to 73	NHL repeat protein	PS50005, TPR TPR repeat profile
	40 to 73	NHL repeat protein	PS50293, TPR_REGION TPR repeat region circular profile
	149 to 179	NHL repeat protein	PS51125, NHL NHL repeat profile
	185 to 223	NHL repeat protein	PS51125, NHL NHL repeat profile
	227 to 270	NHL repeat protein	PS51125, NHL NHL repeat profile
	278 to 316	NHL repeat protein	PS51125, NHL NHL repeat profile
cdsid_YP_005230694.1	1 to 30	Lipoprotein	PS51257, PROKAR_LIPOPROTEIN Prokaryotic membrane lipoprotein lipid attachment site profile

Table 8 Predicted Functional site in total of 13 proteins



Sequence ID	Position	Protein Names	Functional Site
		Species A	
cdsid_YP_005223894.1	5 to 370	Spermidine/putrescine import ATP-binding protein potA family	PS51305, POTA Spermidine/putrescine import ATP-binding protein potA family profile
	6 to 236	ATP-binding cassette, ABC transporter-type domain	PS50893, ABC_TRANSPORTER_2 ATP-binding cassette, ABC transporter-type domain profile
	136 to 150	ATP-binding cassette, ABC transporter-type domain	PS00211, ABC_TRANSPORTER_1 ABC transporters family signature
		Species B	
nil	nil	nil	nil
		Species C	
cdsid_YP_005631875.1	68 to 86	30S Ribosomal protein S9	PS00360, RIBOSOMAL_S9 Ribosomal protein S9 signature
		Species D	
cdsid_NP_218693.1	233 to 546	Apolipoprotein N-acyltransferase	PS50263, CN_HYDROLASE Carbon-nitrogen hydrolase domain profile
cdsid_NP_219333.1	9 to 513	ATP-dependent nuclease, subunit A	PS51198, UVRD_HELICASE_ATP_BIND UvrD-like DNA helicase ATP-binding domain profile
	566 to 850	ATP-dependent nuclease, subunit A	PS51217, UVRD_HELICASE_CTER UvrD-like DNA helicase C-terminal domain profile
		Species E	
nil	nil	nil	nil

Protein functional profile searches resulted in the identification of diverged and common components might mediate interactions between the spirochetes and host cells or the extracellular matrix. It appears possible to tentatively understand the divergent mechanisms underlying their invertebrate pathogenesis and virulence and adaptation to their specific niches.
Comparative analysis of Spirocheate *Treponema pallidum* five subspecies was performed.

1. The parameters include comparison of the genomic and protein function similarity and differences on the basis of genome sequences obtained from NCBI were studied.
2. This study enables the further analysis of the species to understand and grasp the growth, development and impact of research and to research on the pathogenecity activity to overcome the diseases and for treatment and prevention of the causative agent of diseases caused by these organisms.

Finally, our strategy has demonstrated the discriminatory power of computational tools and techniques with Support Vector Machine classification as the sequence based comparative analysis to discriminate proteins and their functions associated within the species of pathogenic microorganisms with high reliability and accuracy.

SUMMARY OF RESEARCH

The aim of the study is to allocate the common and differentiating functional components encoded within the genomes of the five species of *Treponema*. Research review signifies the similarities and differences at morphological level as, *Treponema pallidum* subspecies are morphologically and serologically indistinguishable. This study has demonstrated the discriminatory power of computational tools and techniques with Support Vector Machine classification as the sequence based comparative analysis to discriminate proteins and their functions associated within the species of pathogenic microorganisms with high reliability and accuracy.

FUTURE ISSUES

Sequence variants could be readily used for molecular typing and identification of these Treponema pallidum strains and, with accumulation of additional data from other Treponema pallidum genomes, for epidemiologic applications and clinical discrimination between reinfection or reactivation of diseases. Moreover, the ability to now sequence numerous T.pallidum strains, especially those showing different degrees of virulence, will allow phenotype to be correlated with sequence. This is a significant development for an organism of important public health impact, but for which standard bacterial genetic methods is untenable. We hope that this work can be extended by exploring further sequence properties as well as more diverse organisms, to elucidate the underlying host association and evolutionary mechanisms

DISCLOSURE STATEMENT

There is no financial support for this research work from the funding agency.

ACKNOWLEDGMENTS

We thank our guide for his timely help, giving outstanding ideas and encouragement to finish this research work successfully.

REFERENCES

1. D elcher AL, H armon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 1999, 27, 4636-4641
2. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997, 25, 955-964
3. **Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Research* 1998, 26, 1107-1115**
4. De Castro E., Sigrist C.J.A., Gattiker A., Bulliard V., Langendijk-Genevaux P.S., Gasteiger E., Bairoch A., Hulo N. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res*. 2006, 1, 34(Web Server issue), W362-5
5. Sigrist CJA, De Castro E, Langendijk-Genevaux PS, Le Saux V, Bairoch A, Hulo N. ProRule: a new database containing functional and structural information on PROSITE profiles. *Bioinformatics*, 2005, 21(21), 4060-6
6. Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 2007, 24,1596-1599
7. Sigrist CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N. PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res*. 2010, 38(Database issue), 161-6

Lukashin et al., (1998): In this study, researchers present the analysis of false positive and false negative predictions with the caution that these categories are not precisely defined if the public database annotation is used as a control.

RELATED RESOURCE

1. ArdhaniDwi Lestari, TiniPalupi, Bertha Oktarina, MochammadYuwono, GunawanIndrayanto. *J. Liquid Chromatography & Related Technol.,* 2004, 25(27), 2603-2612
2. Rabert Hartman, Ahmed Abrahim, Andrew Claused, Bing Mao, Louis S. Crocker, ZhihongGe. *J. Liquid Chromatography and Related Technol.,* 2003, 25(26), 2551-2566